INSTITUTE OF NUCLEAR AND PARTICLE PHYSICS

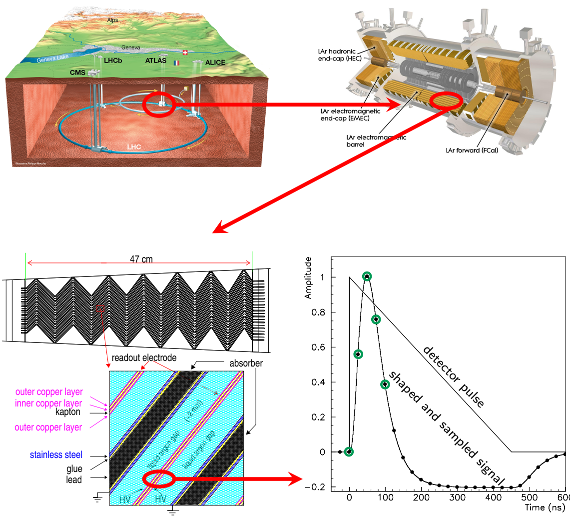
Machine Learning for Real-Time Processing of ATLAS Liquid Argon Calorimeter Signals with FPGAs

Nick Fritzsche
TU Dresden
on behalf of the ATLAS Liquid Argon Calorimeter Group

8 July, 2022

ICHEP 2022 BOLOGNA

The ATLAS Liquid Argon (LAr) Calorimeters

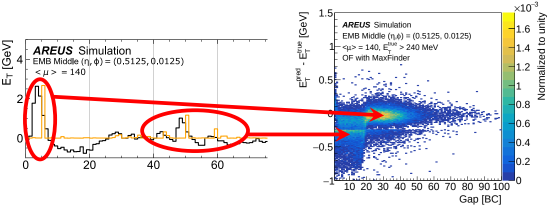


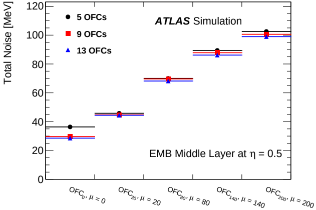
©CERN

1. ATLAS detector at LHC contains **sampling calorimeters** for measurement of energy deposited by electrons, photons and hadronic jets
2. ~ 182k cells
   1. active material: **liquid argon**
   2. absorber: lead, copper, tungsten
3. **Triangular pulse** by ionization is amplified, shaped and digitized at 40 MHz
4. Energy reconstruction with **Optimal Filter (OF)**

$$\underbrace{E(t)}_{\text{Energy}} = \sum_{i=t-N}^{t} \underbrace{a_i}_{\text{Coef}} \cdot \underbrace{x_i}_{\text{Sample}}$$

Signal Processing at High Luminosity LHC (HL-LHC)





CERN-LHCC-2017-018

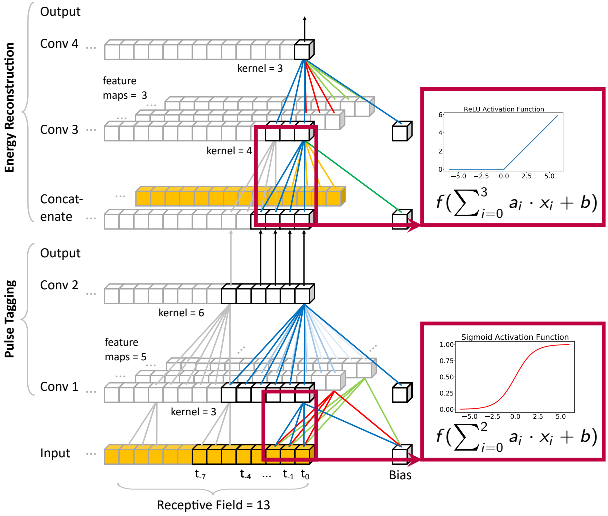
Upgrade Challenges

1. HL-LHC planned to start in 2029 with 7.5x nominal luminosity
   $\longrightarrow$ 140-200 proton-proton collisions per bunch crossing (currently $\sim$ 40)
2. Close-by and overlapping pulses biased or missed
3. Increasing receptive field does not improve OF performance much

$\longrightarrow$ Advanced processing algorithms required

Hardware Trigger

Trigger selects events after $\sim 2\,\mu s$, 150 ns foreseen for energy reconstruction
$\longrightarrow$ Implement algorithms on FPGA for real-time processing
$\longrightarrow$ Short latency and FPGA resources limit complexity of algorithms



Architecture

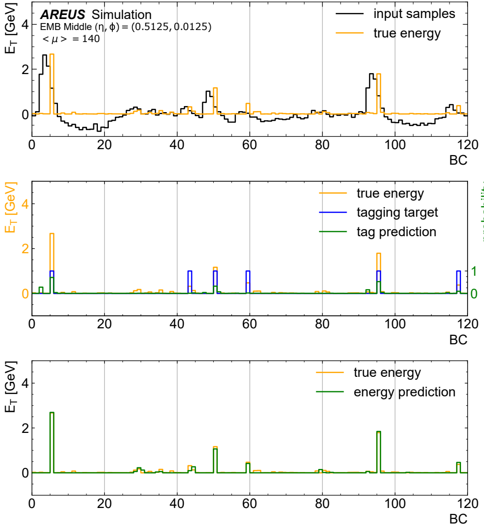
2-tier convolutional network:

1. pulse tagging
2. energy reconstruction

Layer operations

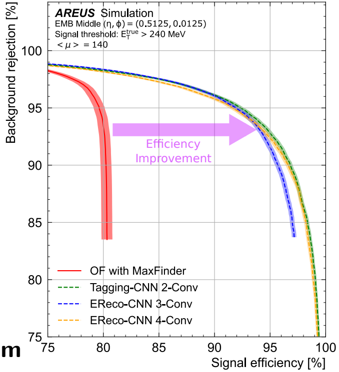
1. Linear combination of previous layer
2. Apply activation function

CNN Training



Training targets

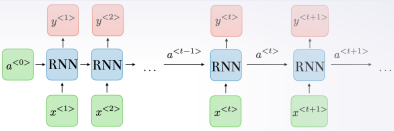
1. Overall: Determine deposited energy out of ADC sequence
2. Pulse tagging: Binary sequence, find energy deposits
3. Energy reconstruction: Value at amplitude $\sim$ energy



$\longrightarrow$ **CNNs outperform legacy OF algorithm**

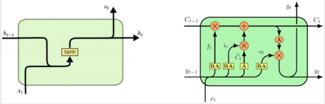
RNN Architectures

Process new input combined with previous state



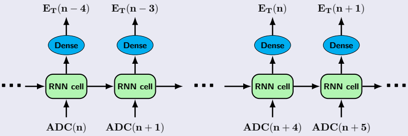
Two internal RNN architectures explored:

1. Long Short-Term Memory (LSTM)
2. Vanilla-RNN, fewer internal dimensions



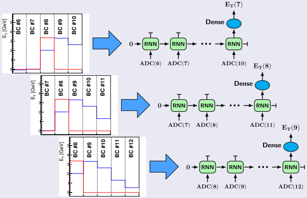
Single Cell

- Long range correction, full signal processed in a stream
- High complexity needed, only LSTM

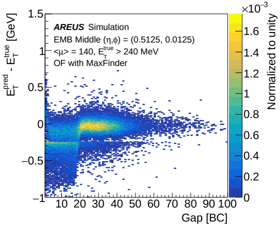


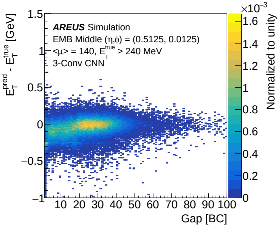
Sliding Window Method

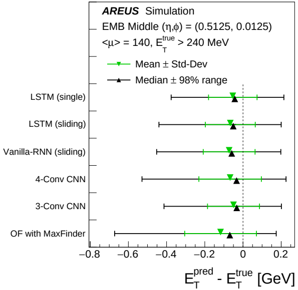
- No long-range correlations, simpler training
- Short range correction only

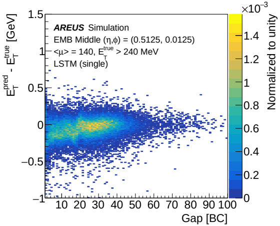


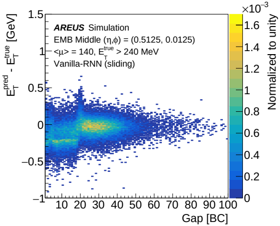
Performance under HL-LHC conditions











Performance

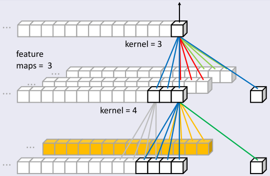
1. All ANNs outperform OF
2. Single LSTM shows best energy resolution and close-by pulse identification
3. CNNs and Vanilla-RNN good compromises between complexity and performance

ANNs on FPGAs

FPGA implementation required for running ANNs on hardware

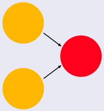
1. Operations mapped to FPGA configurable logic, DSPs, memory, ...
2. Fixed-point arithmetic applied
3. Support time-division multiplexing

CNNs



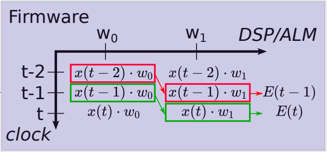
©INTEL

1. CNNs use custom converter from software model to VHDL
2. DSP chain designed for low latency and efficient resource usage



Software

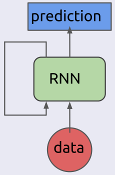
$$E(t-1) = x(t-1) \cdot w_1 + x(t-2) \cdot w_0$$
$$E(t) = x(t) \cdot w_1 + x(t-1) \cdot w_0$$



RNNs implemented in Intel High Level Synthesis (HLS) and VHDL

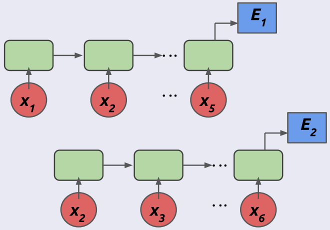
Single Cell & Sliding Window Implementation

Single Cell:



- Single RNN instance on hardware

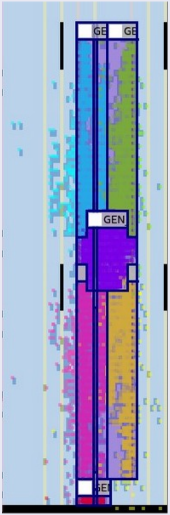
Sliding Window:



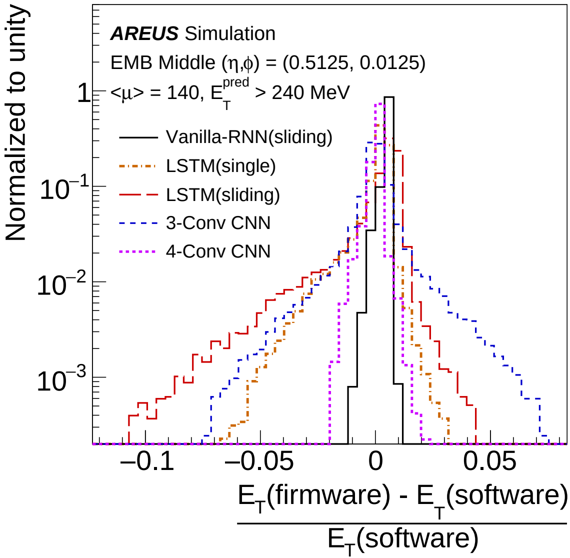
- 5 RNN instances
- Independent pipelined sequences

Placement Constraints

Logic locked region for each cell improves timing



Performance on FPGA



Software model vs firmware implementation

1. Floating-point numbers vs fixed-point calculations
2. Good agreement of FPGA implementation with software
3. Confirmed for CNNs with bit-exact software model of firmware

FPGA Resource Usage

- Process 384 calorimeter cells per FPGA → In total ~ 550 FPGAs needed
- Time-division multiplexing: e.g. FPGA frequency 480 MHz = $12 \cdot 40$ MHz
  → Process 12 cells in pipeline on 1 ANN instance

Single Channel

	3-Conv CNN	4-Conv CNN	Vanilla RNN (sliding)	LSTM (single)	LSTM (sliding)
Frequency F_{max} [MHz]	493	480	641	560	517
Latency clk_{core} cycles	62	58	206	220	363
Resource Usage #DSPs	46 0.8%	42 0.7%	34 0.6%	176 3.1%	738 12.8%
#ALMs	5684 0.6%	5702 0.6%	13115 1.4%	18079 1.9%	69892 7.5%

Time-multiplexed

	3-Conv CNN	4-Conv CNN	Vanilla RNN (HLS)	Vanilla RNN (VHDL)	28× Vanilla RNN (VHDL)
Multiplicity	12	12	10	14	14
Frequency F_{max} [MHz]	487	423	455	587	561
Latency [ns]	125	150	302	121	121
Max. Channels	516	660	370	588	392*
Resource Usage #DSPs	46 0.8%	42 0.7%	152 2.6%	136 2.4%	3808 66.1%
#ALMs	21256 2.3%	16698 1.8%	24433 2.6%	5854 0.6%	164321 17.6%

* For 28 RNNs with 14-fold multiplexing

High latency & resource usage for LSTMs
→ Focus on Vanilla RNN

ALMs shared with other firmware components
→ Optimizations needed

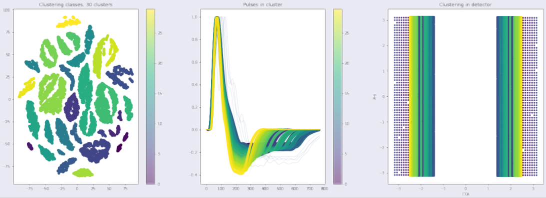
VHDL implementation outperforms HLS

Outlook

Further simulation studies and implementation improvements ongoing

Architecture & Training

- Consider more realistic conditions
  1. Varying pulse shapes
  2. Time shifts to optimal sampling point
  3. LHC bunch train structure
  4. Quantization aware training
  5. **Different detector regions**
- Add new features
  1. Provide timing of detected pulse as output



Firmware & Hardware

1. Optimize firmware implementation
   1. Reduce resource usage
   2. Increase operation frequency
2. Test ANNs on Stratix-10 hardware
   - Integrate ANNs into higher level LAr signal processor firmware

Summary

1. Advanced signal processing algorithms required for ATLAS LAr energy reconstruction under HL-LHC conditions
    - Two machine learning based approaches: CNNs and RNNs
2. Various ANN algorithms studied
    - CNNs and RNNs outperform legacy Optimal Filter algorithm
3. FPGA implementation for real-time processing with high bandwidth developed
    1. CNNs: VHDL implementation
    2. RNNs: High level synthesis and VHDL implementation
4. Promising results of firmware evaluation
    1. Good reproduction of Keras results with firmware simulation
    2. Optimizations ongoing to improve resource usage and latency

$\longrightarrow$ CNNs/RNNs show great potential to improve energy reconstruction of ATLAS LAr calorimeter system under HL-LHC conditions

Ref. "Artificial Neural Networks on FPGAs for Real-Time Energy Reconstruction of the ATLAS LAr Calorimeters"
Aad, G. et al., Comput Softw Big Sci 5, 19 (2021)

Ref. "Energy reconstruction in a liquid argon calorimeter cell using convolutional neural networks"
Polson, L. et al., JINST 17, P01002 (2022)